

1 **CLAIMS**

2
3 1. A method of identifying one or more portions of a document, the
4 method comprising:

5 identifying a plurality of visual blocks in the document;
6 detecting one or more separators between the visual blocks of the plurality
7 of visual blocks; and
8 constructing, based at least in part on the plurality of visual blocks and the
9 one or more separators, a content structure for the document, wherein the content
10 structure identifies the different visual blocks as different portions of semantic
11 content of the document.

12
13 2. A method as recited in claim 1, wherein the document comprises a
14 web page.

15
16 3. A method as recited in claim 1, wherein the document is described by
17 a tree structure having a plurality of nodes, and wherein identifying the plurality of
18 visual blocks in the document comprises:

19 identifying a group of candidate nodes of the plurality of nodes;
20 for each node in the group of candidate nodes:
21 determining whether the node can be divided, and
22 if the node cannot be divided, then identifying the node as
23 representing a visual block.

1 4. A method as recited in claim 3, wherein if the node cannot be
2 divided, then setting a degree of coherence for the visual block represented by the
3 node.

4
5 5. A method as recited in claim 3, wherein if the node cannot be
6 divided, then removing the node from the group of candidate nodes.

7
8 6. A method as recited in claim 3, wherein determining whether the
9 node can be divided comprises determining that the node can be divided if the
10 node has a child node with <HR> HyperText Markup Language (HTML) tag.

11
12 7. A method as recited in claim 3, wherein determining whether the
13 node can be divided comprises determining that the node can be divided if a
14 background color of the node is different from a background color of a child of the
15 node.

16
17 8. A method as recited in claim 3, further comprising checking whether
18 the node has a child having a width and height greater than zero, and if the node
19 has no child having a width and height greater than zero then removing the node
20 from the group of candidate nodes.

1 **9.** A method as recited in claim 3, wherein determining whether the
2 node can be divided comprises determining that the node can be divided if a size
3 of the node is at least a threshold amount greater than a sum of sizes of children
4 nodes of the node.

5
6 **10.** A method as recited in claim 3, wherein determining whether the
7 node can be divided comprises determining that the node can be divided if the
8 node has multiple successive children nodes each having a
 HyperText
9 Markup Language (HTML) tag.

10
11 **11.** A method as recited in claim 1, wherein the document is described
12 by a tree structure having a plurality of nodes, and wherein identifying the
13 plurality of visual blocks in the document comprises identifying different visual
14 blocks based at least in part on HyperText Markup Language (HTML) tags of the
15 plurality of nodes.

16
17 **12.** A method as recited in claim 1, wherein the document is described
18 by a tree structure having a plurality of nodes, and wherein identifying the
19 plurality of visual blocks in the document comprises identifying different visual
20 blocks based at least in part on background colors of the plurality of nodes.

1 **13.** A method as recited in claim 1, wherein the document is described
2 by a tree structure having a plurality of nodes, and wherein identifying the
3 plurality of visual blocks in the document comprises identifying different visual
4 blocks based at least in part on whether the plurality of nodes include text and the
5 sizes of the plurality of nodes.

6
7 **14.** A method as recited in claim 1, wherein detecting the one or more
8 separators comprises:

9 detecting one or more horizontal separators between the visual blocks; and
10 detecting one or more vertical separators between the visual blocks.

11
12 **15.** A method as recited in claim 1, wherein detecting the one or more
13 separators comprises:

14 initializing a separator list that includes one or more possible separators
15 between the visual blocks;

16 analyzing, for each of the visual blocks, whether the visual block overlaps a
17 separator of the separator list, and if so how the visual block overlaps the
18 separator; and

19 determining how to treat the separator based on whether the visual block
20 overlaps the separator, and if so how the visual block overlaps the separator.

21
22 **16.** A method as recited in claim 15, further comprising determining to
23 split the separator into multiple separators if the visual block is contained in the
24 separator.
25

1 **17.** A method as recited in claim 15, further comprising determining, if
2 the visual block crosses the separator, to modify parameters of the separator so
3 that the visual block no longer crosses the separator.
4

5 **18.** A method as recited in claim 17, wherein the modification
6 comprises reducing the height of the separator if the separator is a horizontal
7 separator.
8

9 **19.** A method as recited in claim 17, wherein the modification
10 comprises reducing the width of the separator if the separator is a vertical
11 separator.
12

13 **20.** A method as recited in claim 15, further comprising determining to
14 remove the separator from the separator list if the visual block covers the
15 separator.
16

17 **21.** A method as recited in claim 1, further comprising assigning, to
18 each of the one or more separators, a weight based on characteristics of visual
19 blocks on either side of the separator.
20

21 **22.** A method as recited in claim 21, wherein assigning the weight
22 comprises assigning the weight based on a distance between two visual blocks on
23 either side of the separator.
24
25

1 **23.** A method as recited in claim 21, wherein assigning the weight
2 comprises assigning the weight based on whether the separator is at a same
3 position as an <HR> HTML tag.

4
5 **24.** A method as recited in claim 21, wherein assigning the weight
6 comprises assigning the weight based on a font size used in two visual blocks on
7 either side of the separator.

8
9 **25.** A method as recited in claim 21, wherein assigning the weight
10 comprises assigning the weight based on a background color used in two visual
11 blocks on either side of the separator.

12
13 **26.** A method as recited in claim 1, further comprising:
14 checking whether each of the plurality of visual blocks satisfies a degree of
15 coherence threshold; and
16 for each of the plurality of visual blocks that does not satisfy the degree of
17 coherence threshold, identifying a new plurality of visual blocks in the visual
18 block, and repeating the detecting and constructing using the new plurality of
19 visual blocks.

20
21 **27.** A method as recited in claim 1, wherein constructing the content
22 structure comprises:
23 generating one or more virtual blocks based on the plurality of visual
24 blocks; and
25 including, in the content structure, the one or more virtual blocks.

1
2 **28.** A method as recited in claim 27, wherein generating the one or more
3 virtual blocks comprises generating the one or more virtual blocks by combining
4 two visual blocks of the plurality of visual blocks.

5
6 **29.** A method as recited in claim 27, further comprising:
7 determining a degree of coherence value for each of the one or more virtual
8 blocks.

9
10 **30.** A method as recited in claim 29, wherein determining the degree of
11 coherence value for a virtual block comprises determining the degree of coherence
12 value for the virtual block based at least in part on a weight of a separator between
13 two visual blocks used to generate the virtual block.

14
15 **31.** One or more computer readable media having stored thereon a
16 plurality of instructions that, when executed by one or more processors of a
17 device, causes the one or more processors to:

18 identify visual blocks in a document;
19 detect visual separators between the visual blocks; and
20 construct, based at least in part on the visual blocks and the visual
21 separators, a content structure for the document that identifies regions of the
22 document that represent semantic content of the document.

1 **32.** One or more computer readable media as recited in claim 31,
2 wherein the document is described by a tree structure having a plurality of nodes,
3 and wherein the instructions that cause the one or more processors to identify
4 visual blocks in the document comprise instructions that cause the one or more
5 processors to:

6 identify a group of candidate nodes of the plurality of nodes;

7 for each node in the group of candidate nodes:

8 determine whether the node can be divided, and

9 if the node cannot be divided, then identify the node as representing
10 a visual block.

11
12 **33.** One or more computer readable media as recited in claim 31,
13 wherein the instructions that cause the one or more processors to detect visual
14 separators comprise instructions that cause the one or more processors to:

15 detect one or more horizontal separators between the visual blocks; and

16 detect one or more vertical separators between the visual blocks.

17
18 **34.** One or more computer readable media as recited in claim 31,
19 wherein the instructions that cause the one or more processors to detect visual
20 separators comprise instructions that cause the one or more processors to:

21 initialize a separator list that includes one or more possible visual
22 separators between the visual blocks;

23 analyze, for each of the visual blocks, whether the visual block overlaps a
24 separator of the separator list, and if so how the visual block overlaps the
25 separator; and

1 determine how to treat the separator based on whether the visual block
2 overlaps the separator, and if so how the visual block overlaps the separator.

3
4 **35.** One or more computer readable media as recited in claim 31,
5 wherein the instructions further cause the one or more processors to:

6 check whether each of the visual blocks satisfies a degree of coherence
7 threshold; and

8 for each of the visual blocks that does not satisfy the degree of coherence
9 threshold, identify new visual blocks in the visual block, and repeat the detection
10 and construction using the new visual blocks.

11
12 **36.** A method of searching a plurality of documents, the method
13 comprising:

14 receiving query criteria corresponding to a query;

15 accessing a plurality of blocks corresponding to the plurality of documents,
16 wherein different blocks of the plurality of blocks correspond to different
17 documents of the plurality of documents, wherein the plurality of blocks have
18 been obtained by visually segmenting each of the plurality of documents;

19 generating rankings for one or more of the plurality of blocks based at least
20 in part on how well the blocks match the query criteria;

21 generating rankings for one or more of the plurality of documents, wherein
22 the ranking of each of the plurality of documents is based at least in part on the
23 rankings of the multiple blocks corresponding to the document; and

24 returning an indication of at least one of the one or more ranked documents.
25

1 **37.** A method as recited in claim 36, wherein each of the plurality of
2 documents comprises a web page.

3
4 **38.** A method as recited in claim 36, wherein generating the ranking for
5 one of the plurality of documents comprises:

6 identifying the rankings of each of the multiple blocks corresponding to the
7 one document;

8 selecting, as the ranking for the one document, the highest ranking of the
9 identified rankings.

10
11 **39.** A method as recited in claim 36, wherein generating the ranking for
12 one of the plurality of documents comprises:

13 identifying the rankings of each of the multiple blocks corresponding to the
14 one document; and

15 combining the identified rankings to form a ranking for the one document.

16
17 **40.** A method as recited in claim 39, wherein the combining comprises
18 averaging the identified rankings.

19
20 **41.** A method as recited in claim 36, wherein the visually segmenting a
21 document comprises:

22 identifying a plurality of visual blocks in the document;

23 detecting one or more separators between the visual blocks of the plurality
24 of visual blocks; and

1 constructing, based at least in part on the plurality of visual blocks and the
2 one or more separators, a content structure for the document, wherein the content
3 structure identifies the different visual blocks as different portions of semantic
4 content of the document, and wherein the different visual blocks are the blocks of
5 the plurality of blocks that correspond to the document.

6
7 **42.** A method as recited in claim 41, wherein the document is described
8 by a tree structure having a plurality of nodes, and wherein identifying the
9 plurality of visual blocks in the document comprises:

10 identifying a group of candidate nodes of the plurality of nodes;

11 for each node in the group of candidate nodes:

12 determining whether the node can be divided, and

13 if the node cannot be divided, then identifying the node as
14 representing a visual block.

15
16 **43.** One or more computer readable media having stored thereon a
17 plurality of instructions that, when executed by one or more processors of a
18 device, causes the one or more processors to:

19 receive a query including one or more search terms;

20 rank a plurality of blocks based on how well the plurality of blocks matches
21 the one or more search terms, wherein each of the plurality of blocks is part of one
22 document of a plurality of documents, and wherein each of the plurality of blocks
23 is obtained by visual segmentation of one of the plurality of documents;

24 for each of the plurality of documents, rank the document based at least in
25 part on the rankings of the blocks that are part of the document; and

1 return, in response to the query, an indication of the rankings of one or
2 more of the plurality of documents.

3
4 **44.** One or more computer readable media as recited in claim 43,
5 wherein the instructions that cause the one or more processors to rank the
6 document comprise instructions that cause the one or more processors to:

7 identify the ranking for each block that is part of the document;
8 select, as the ranking for the document, the highest ranking of the identified
9 rankings.

10
11 **45.** One or more computer readable media as recited in claim 43,
12 wherein the instructions that cause the one or more processors to rank the
13 document comprise instructions that cause the one or more processors to:

14 identify the ranking for each block that is part of the document;
15 combine the rankings for each block to generate a ranking for the
16 document.

17
18 **46.** One or more computer readable media as recited in claim 43,
19 wherein the visual segmentation of a document comprises:

20 identifying a plurality of visual blocks in the document;
21 detecting one or more separators between the visual blocks of the plurality
22 of visual blocks; and
23 constructing, based at least in part on the plurality of visual blocks and the
24 one or more separators, a content structure for the document, wherein the content
25 structure identifies the different visual blocks as different portions of semantic

1 content of the document, and wherein the different visual blocks are the blocks of
2 the plurality of blocks that are part of the document.

3
4 **47.** One or more computer readable media as recited in claim 46,
5 wherein the document is described by a tree structure having a plurality of nodes,
6 and wherein identifying the plurality of visual blocks in the document comprises:

7 identifying a group of candidate nodes of the plurality of nodes;

8 for each node in the group of candidate nodes:

9 determining whether the node can be divided, and

10 if the node cannot be divided, then identifying the node as
11 representing a visual block.

12
13 **48.** A method of searching a plurality of web pages, the method
14 comprising:

15 receiving a request to search the plurality of web pages;

16 generating a first set of rankings for a subset of the plurality of web pages
17 based on the request;

18 generating a second set of rankings for the subset of web pages by visually
19 segmenting each web page in the subset of web pages; and

20 obtaining, based at least in part on the second set of rankings, a final set of
21 rankings for the subset of web pages.

22
23 **49.** A method as recited in claim 48, wherein obtaining the final set of
24 rankings comprises using, as the final set of rankings, the second set of rankings.
25

1 **50.** A method as recited in claim 48, wherein obtaining the final set of
2 rankings comprises selecting, as the final ranking for a web page, the higher
3 ranking of the ranking of the web page in the first set and the ranking of the web
4 page in the second set.

5
6 **51.** A method as recited in claim 48, wherein obtaining the final set of
7 rankings comprises averaging, to obtain the final ranking for a web page, the
8 ranking of the web page in the first set and the ranking of the web page in the
9 second set.

10
11 **52.** A method as recited in claim 48, wherein visually segmenting a web
12 page comprises:

13 identifying a plurality of visual blocks in the web page;
14 detecting one or more separators between the visual blocks of the plurality
15 of visual blocks; and
16 constructing, based at least in part on the plurality of visual blocks and the
17 one or more separators, a content structure for the web page, wherein the content
18 structure identifies the different visual blocks as different portions of semantic
19 content of the web page.

20
21 **53.** A method as recited in claim 52, wherein the web page is described
22 by a tree structure having a plurality of nodes, and wherein identifying the
23 plurality of visual blocks in the web page comprises:

24 identifying a group of candidate nodes of the plurality of nodes;
25 for each node in the group of candidate nodes:

1 determining whether the node can be divided, and
2 if the node cannot be divided, then identifying the node as
3 representing a visual block.
4

5 **54.** One or more computer readable media having stored thereon a
6 plurality of instructions that, when executed by one or more processors of a
7 device, causes the one or more processors to:

8 generate first rankings for a plurality of documents based on how well the
9 plurality of documents match search criteria;

10 generate second rankings for the plurality of documents by visually
11 segmenting each of the plurality of documents; and

12 generate final rankings for the plurality of documents based at least in part
13 on the second rankings.
14

15 **55.** One or more computer readable media as recited in claim 54,
16 wherein the instructions that cause the one or more processors to generate final
17 rankings comprise instructions that cause the one or more processors to use, as the
18 final rankings, the second rankings.
19
20
21
22
23
24
25

1 **56.** One or more computer readable media as recited in claim 54,
2 wherein the instructions that cause the one or more processors to generate final
3 rankings comprise instructions that cause the one or more processors to select, as a
4 final ranking for a document of the plurality of documents, whichever ranking of
5 the first ranking for the document and the second ranking of the document is
6 higher.

7
8 **57.** One or more computer readable media as recited in claim 54,
9 wherein the instructions that cause the one or more processors to generate final
10 rankings comprise instructions that cause the one or more processors to generate a
11 final ranking for a document of the plurality of documents by averaging the first
12 ranking of the document and the second ranking of the document.

13
14 **58.** One or more computer readable media as recited in claim 54,
15 wherein the instructions that cause the one or more processors to visually segment
16 a document comprise instructions that cause the one or more processors to:

17 identify a plurality of visual blocks in the document;

18 detect one or more separators between the visual blocks of the plurality of
19 visual blocks; and

20 construct, based at least in part on the plurality of visual blocks and the one
21 or more separators, a content structure for the document, wherein the content
22 structure identifies the different visual blocks as different portions of semantic
23 content of the document.

24

25

1 **59.** One or more computer readable media as recited in claim 58,
2 wherein the document is described by a tree structure having a plurality of nodes,
3 and wherein the instructions that cause the one or more processors to identify the
4 plurality of visual blocks in the document comprise instructions that cause the one
5 or more processors to:

6 identify a group of candidate nodes of the plurality of nodes;

7 for each node in the group of candidate nodes:

8 determine whether the node can be divided, and

9 if the node cannot be divided, then identify the node as representing
10 a visual block.

11
12 **60.** A method of searching a plurality of documents, the method
13 comprising:

14 receiving a request to search the plurality of documents, wherein the
15 request includes query criteria;

16 identifying a subset of the plurality of documents based on the query
17 criteria;

18 identifying, for each of the subset of documents, a plurality of blocks by
19 visually segmenting the document;

20 expanding, based on the content of the plurality of blocks, the query
21 criteria; and

22 identifying a second subset of the plurality of documents based on the
23 expanded query criteria.
24
25

1 **61.** A method as recited in claim 60, returning, in response to the
2 request, identifiers of the second subset of documents.

3
4 **62.** A method as recited in claim 60, ranking each document of the
5 second subset of the plurality of documents; and
6 returning, in response to the request, identifiers of the second subset of
7 documents and an indication of the ranking of each document of the second subset
8 of documents.

9
10 **63.** A method as recited in claim 60, wherein the visually segmenting
11 the document comprises:

12 identifying a plurality of visual blocks in the document;
13 detecting one or more separators between the visual blocks of the plurality
14 of visual blocks; and
15 constructing, based at least in part on the plurality of visual blocks and the
16 one or more separators, a content structure for the document, wherein the content
17 structure identifies the different visual blocks as different portions of semantic
18 content of the document, and wherein the different visual blocks are the plurality
19 of blocks for the document.

20
21 **64.** A method as recited in claim 63, wherein the document is described
22 by a tree structure having a plurality of nodes, and wherein identifying the
23 plurality of visual blocks in the document comprises:

24 identifying a group of candidate nodes of the plurality of nodes;
25 for each node in the group of candidate nodes:

1 determining whether the node can be divided, and
2 if the node cannot be divided, then identifying the node as
3 representing a visual block.
4

5 **65.** One or more computer readable media having stored thereon a
6 plurality of instructions that, when executed by one or more processors of a
7 device, causes the one or more processors to:

8 receive one or more search terms;

9 identify a plurality of documents that satisfy the one or more search terms;

10 perform vision-based document segmentation on each of the plurality of
11 documents to identify blocks of each of the plurality of documents;

12 generate a rank for each of the identified blocks based on how well the
13 block matches the one or more search terms;

14 derive one or more expansion terms from one or more of the identified
15 blocks; and

16 identify another plurality of documents that satisfy the one or more search
17 terms and the expansion terms.
18

19 **66.** One or more computer readable media as recited in claim 65,
20 wherein the instructions that cause the one or more processors to derive the one or
21 more expansion terms cause the one or more processors to derive the one or more
22 expansion terms from a group of top-ranked identified blocks.
23
24
25

1 **67.** One or more computer readable media as recited in claim 65,
2 wherein the instructions that cause the one or more processors to perform vision-
3 based document segmentation comprise instructions that cause the one or more
4 processors to:

5 identify a plurality of visual blocks in the document;

6 detect one or more separators between the visual blocks of the plurality of
7 visual blocks; and

8 construct, based at least in part on the plurality of visual blocks and the one
9 or more separators, a content structure for the document, wherein the content
10 structure identifies the different visual blocks as different portions of semantic
11 content of the document, and wherein the different visual blocks are the blocks of
12 the document.

13
14 **68.** A system comprising:

15 a visual block extractor to extract visual blocks from a document;

16 a visual separator detector coupled to receive the extracted visual blocks
17 and detect, based on the extracted visual blocks, one or more visual separators
18 between the extracted visual blocks; and

19 a content structure constructor coupled to receive the extracted visual
20 blocks and the detected visual separators, and to use the extracted visual blocks
21 and the detected visual separators to construct a content structure for the
22 document.

1 **69.** A system as recited in claim 68, further comprising:
2 a document retrieval module to retrieve documents from a plurality of
3 documents based at least in part on the content structure constructed for one or
4 more of the plurality of documents.

5
6 **70.** A system as recited in claim 68, wherein the document is described
7 by a tree structure having a plurality of nodes, and wherein the visual block
8 extractor is to extract visual blocks from the document by:

9 identifying a group of candidate nodes of the plurality of nodes;

10 for each node in the group of candidate nodes:

11 determining whether the node can be divided, and

12 if the node cannot be divided, then identifying the node as
13 representing a visual block.

14
15 **71.** A system as recited in claim 68, wherein the visual separator
16 detector is to detect one or more horizontal separators between the visual blocks
17 and one or more vertical separators between the visual blocks.

18
19 **72.** A system as recited in claim 68, wherein the visual separator
20 detector is to detect the one or more separators by:

21 initializing a separator list that includes one or more possible separators
22 between the visual blocks;

23 analyzing, for each of the visual blocks, whether the visual block overlaps a
24 separator of the separator list, and if so how the visual block overlaps the
25 separator; and

1 determining how to treat the separator based on whether the visual block
2 overlaps the separator, and if so how the visual block overlaps the separator.

3
4 **73.** A system as recited in claim 68, wherein the content structure
5 constructor is further to:

6 check whether each of the plurality of visual blocks satisfies a degree of
7 coherence threshold; and

8 for each of the plurality of visual blocks that does not satisfy the degree of
9 coherence threshold, return the visual block to the visual block extractor to have a
10 new plurality of visual blocks extracted from the visual block, and further to have
11 the visual separator detector detect one or more visual separators using the new
12 plurality of visual blocks.

13
14 **74.** A system comprising:

15 means for identifying a plurality of visual blocks in the document;

16 means for detecting one or more separators between the visual blocks of the
17 plurality of visual blocks; and

18 means for constructing, based at least in part on the plurality of visual
19 blocks and the one or more separators, a content structure for the document,
20 wherein the content structure identifies the different visual blocks as different
21 portions of semantic content of the document.

1 75. A system as recited in claim 74, wherein the document is described
2 by a tree structure having a plurality of nodes, and wherein the means for
3 identifying the plurality of visual blocks in the document comprises:

4 means for identifying a group of candidate nodes of the plurality of nodes;

5 for each node in the group of candidate nodes:

6 means for determining whether the node can be divided, and

7 means for identifying, if the node cannot be divided, the node as
8 representing a visual block.